CONFIDENTIAL DATA IDENTIFICATION USING DATA MINING TECHNIQUES IN DATA LEAKAGE PREVENTION SYSTEM

Peneti Subhashini and B Padmaja Rani

Department of Computer Science Engineering,
Jawaharlal Nehru Technological University,
Hyderabad, Telangana, India-500085.

ABSTRACT

Data leakage means sending confidential data to an unauthorized person. Nowadays, identifying confidential data is a big challenge for the organizations. We developed a system by using data mining techniques, which identifies confidential data of an organization. First, we create clusters for the training data set. Next, identify confidential terms and context terms for each cluster. Finally, based on the confidential terms and context terms, the confidentiality level of the detected document calculated in terms of score. If the score of the detected document beyond a predefined threshold, then the document is blocked and marked as a confidential.

KEYWORDS

Data leakage, Data mining techniques, Confidential terms, Context terms, Threshold, Confidentiality level.

1. INTRODUCTION

Now day's data leakage is the biggest challenge in all the types of public and private organizations. For the protection of confidential data, data leakage prevention system (DLP) is used by many organizations [1].Confidential information such as customer data, employee data and business secrets which are the main assets of an organization. Confidential information is crucial for the organization's staff and partners to perform their business tasks [2]. Data leakage prevention is the category of solution, which helps an organization to apply controls for preventing the unwanted or malicious leakage of confidential information to unauthorized entities in or outside the organization [3].

Data leakage prevention systems use confidential terms and data identification methods for controlling data leakages in the organization. First, DLP system identifies which documents are confidential documents and non-confidential documents. In this paper, the confidential data identification method classifies the organization documents into confidential and non-confidential. Any employee in the organization wants to send any document to an outside; first the sending document is compared with existing confidential and non-confidential documents of

an organization. If the sending document (whole document or some portions) is matched with any one of the existing confidential documents that is blocked by the DLP.

2. RELATED WORK

In this section we provide review on different confidential data identification methods in Data leakage prevention. Existing confidential data identification methods can be divided into two categories: Content based and Behaviour based methods. The content based method includes rule-based and classifier based approaches.

In content-based approach, directly focuses on data values. Data values contain the use of confidential words, regular expression, text classification and information retrieval. In rule-based approach a group of predefined polices is outlined with respect to words and terms that will seem within the document. These rules determine the confidentiality level of the document. Rule-based approaches suffer from a high rate of false-positives; according to [12].Organizations that uses rule-based approaches usually defined a high threshold for confidential content in documents, as a result of high false-positive rate, rule-based approaches are not used these days [13,14,15,16].

In classifier based approach various classifications and machine learning techniques are used [17, 18], such as SVM (Support Vector Machine) and Naïve Bayes [19, 20, 21]. In this approach the content of the document is represented as vectors [22]. Vectors are generated using terms and their frequencies of the documents. These vectors classifies, whether the documents are confidential or not. Small portions of confidential data are not identified by this approach.
The behaviour-based approach focus on identifying anomalies in behaviour [23]. These anomalies can be tracked in the communication, in and out of the organization, or the analysis of past and current communication.

In this paper, implemented a method in contrast to the above methods. Data mining approaches, confidential word matching approaches are the confidential concepts of this method. In the first step, we used clustering concept to group together a document of similar content. In a second step, extract the confidential terms content of each cluster using language modelling technique. Finally, determine whether a tested document is confidential or not.

3. CONFIDENTIAL DATA IDENTIFICATION METHOD

The projected method consists of two phases. Training phase and detection phase. During the training phase, organization documents are grouped together with the help of clustering and language modelling techniques. During the detection phase confidential score calculated for the inspected document. If the score exceeds the predefined threshold value then that document marked as a confidential. Section 3.1 and section 3.2 provides detailed description about training phase and detection phase.

3.1 Training Phase

The objective of this phase is to representation of confidential content of a document. This representation should include not only confidential words and terms, but also the context in

which they appear [4].The training phase requires two sets of documents as an input. C is the one input, which contains a set of confidential documents. The second input is N, is a set of non-confidential documents. For every document, tokenization [5], remove stop-words next apply stemming algorithm and finally transforming them into vectors of weighted terms [6]. For clustering we used K-means unsupervised algorithm with cosine measure as the distance function. The following is the pictorial representation for training phase [7].

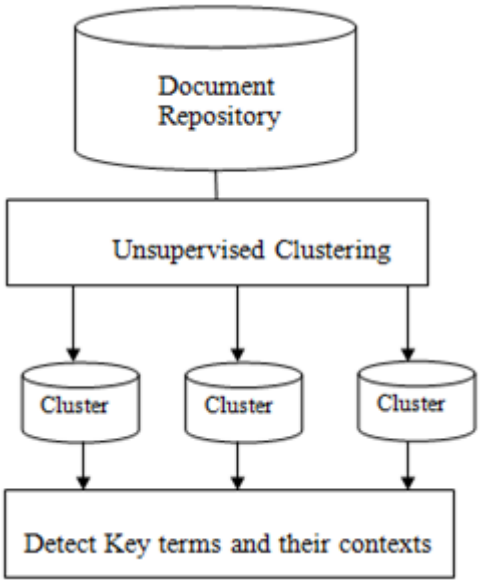

Figure1.Training Phase

The training phase consist of two steps [8]

1. Key terms detection
2. Context terms detection

3.1.1 Key terms detection

Key terms detection is the important step for this method. The key terms have two purposes:

a) They serve as initial indicators of confidential content.
b) They serve of the axis around which the context terms are generated**.**

Without a relevant and a robust set of key terms, the method presented in this paper is not likely to succeed. The key term detection process is based on a technique called language modelling [8].This technique represents the value of a term (or a sequence of terms) in terms of probability.

The detection of the key terms is done as follows [7]:

- For each cluster create a language model separately for confidential (C) and non-confidential documents (N), which denote lm_c and lm_n accordingly.

$$lm\left(\frac{t}{d}\right) = \frac{tf_t}{N_{(t,d)}} \quad (1)$$

tf indicates term frequency and N total no of terms in the document.

- For each term *t* in lm_c calculate its score using the following formula

$$score(t) = \frac{lm_c(t)}{lm_n(t)} \quad (2)$$

- The score assigned to term t reflects how much they more likely appear in a relevant document(C) than in a non-relevant (N) one.

- Only the terms whose assigned score is greater than 1 are used in order to work out the confidentiality of documents throughout the detection phase.

- In order to correct for inaccuracies that may arise due to data sparseness, apply the Dirichlet smoothing technique [9].

- Smoothing, considered a confidential element in improving the performance of language model [10].

3.1.2 Context terms detection

For each confidential term identify context terms. The context terms serve three purposes [7]:

a) They serve as validators, enabling us to work out whether or not the detected confidential terms are actually indicative of relevant content.
b) They allow us to quantify the degree of significance (many relevant context terms=higher relevance for a confidential term)
c) They enable us to see whether or not the detected key terms are connected by analyzing their shared contexts.

The process of detecting the context terms for a single key term- (key $_{term}$) is as follows [8]:

- Find all the instances of key_{term} every in confidential terms and non-confidential terms document.
- For each instance, extract the terms around $key_{term,}$ using a window of size X ($\frac{x}{2}$ terms before the position of the term and $\frac{x}{2}$ terms after it). Every term during this text excerpt are going to be measured as a possible context term. The size of the sliding window is denoted using a parameter referred to as *context span.*
- Each such excerpt will now be considered as a document. Excerpts from confidential terms and non-confidential terms will be denoted as $d_{\text{confidential t}}$ and $d_{\text{non-confidential}}$ documents.
- Create language model for each context term both in confidential and non-confidential.

$$lm(context) = \frac{no\ of\ documents\ contain\ context\ term}{no\ of\ documnets\ contain\ confidential\ term} \quad (3)$$

- Calculate the probability of each context term appears near the key term key_{term} both in the confidential terms and non-confidential terms documents.

- Finally, calculate the score of each possible context term using the following formula.

$$score(t_{context}) = lm_{d_{confidential}}(t_{context}) - lm_{d_{non-confidential}}(t_{context}) \quad (4)$$

- If the score of a context term higher than a predefined threshold value, then that term considered as a context of a confidential term.

3.2 Detection Phase

The Detection phase detects the confidentiality level of the inspected document. Two types of challenges are discussed in the detection phase

1. Detection of whole confidential documents.
2. Detecting little parts of confidential text embedded in an exceedingly larger non-confidential text.

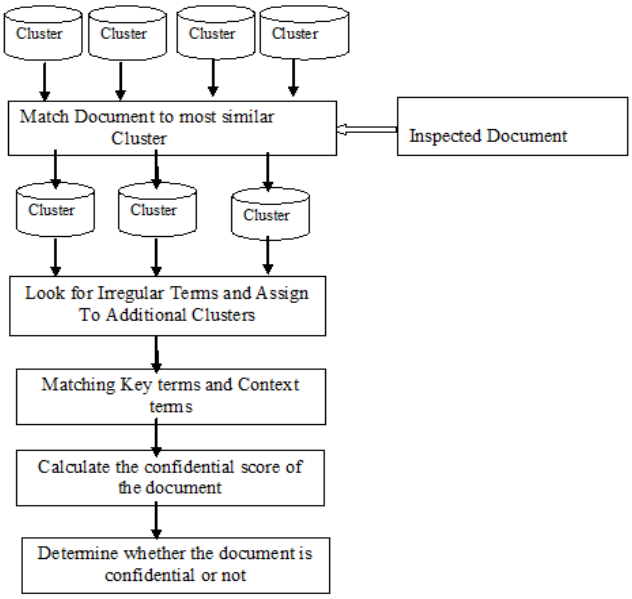

Figure 2. Detection Phase

The inspected document is transformed into a vector (after stemming and stop-words removal) and finds the similar clusters. This is done using the cosine distance measure [6].All the clusters whose similarity are above a predefined threshold are selected.

This similar cluster methodology is working well for detection of whole documents but it is insufficient for detection confidential terms sections embedded in non-confidential terms documents. To solve this problem, we seem to be for irregular terms with in the inspected document, i.e., terms that are most unlikely to look within the document considering the clusters to that it absolutely was assigned. In order to identify these terms, first create language model for both the tested document and the clusters it was assigned to, find score for each term in the inspected document using the following formula

$$\forall t \in D, t_{score} = \frac{p(t/_D)}{p(t/_c)} \quad (5)$$

Where t is that the term within the inspected document D, C is one among the chosen clusters. The upper the score of a term, the less likely it‘s to belong to the cluster. If the term t score is greater than the predefined threshold then that term is considered as an irregular term. Once the terms (irregular) are found they are matched against the confidential terms of every cluster. If match found the corresponding cluster is added to the list of candidate clusters (similar clusters) that the confidentiality score of the document is calculated.

Identify all the confidential and context term that appear both in the inspected document and assigned cluster. In order to discover confidential terms, the inspected document undergoes stemming and stops word elimination and is then matched against the confidential terms represented within the cluster. If the confidential term found, then its corresponding context terms are identified within the cluster.

The confidential terms score of the inspected document for every cluster is calculated by summation of scores of all the confidential terms that met the below criteria.

Table 1.The required number and minimum score of a context terms for every score of the confidential term.

Confidential score of a term	Required no of context terms	Min score of context score
1<score<3	10	80
3<=score<7	9	70
7<=score<10	7	60
10<=score<15	6	50
15<=score<30	5	30
30<=score<45	3	20
45<=score	2	0

If the inspected document confidential terms scores for a cluster above the threshold, then that document marked as a confidential for that cluster and is blocked.

4. EVALUATION

We evaluate the performance of our method on Reuter's news articles dataset.

4.1 Reuters news articles

We had generated one dataset that is compiled from news articles collected from Reuter's news [11]. This dataset consist of 21578 documents. The Reuters 21578 collection is distributed in twenty two files, each of the first twenty one files (reut2-000.sgm-reut2-020.sgm) contains a thousand documents, while the last reut2-021.sgm contains 578 documents. Reuters 21578 divided into 5 categories. For the purpose of the results evaluation, economics category was chosen as the dataset. Economics contains 16 categories, of which trade category chosen as a confidential purpose and remaining 15 categories were chosen as a non-confidential purpose. Confidential documents are 350(200 for training, 100 for detection) and non-confidential are 750(550 for training, 200 for detection).

Table 2.Out of 200 non-confidential documents, the no of non-confidential documents is detected as a confidential for corresponding confidential score and similarity threshold.

Score/threshold	0.2	0.1	0.05
200	20	120	180
150	80	187	197
100	126	197	200

We evaluated our method performance by using True Positive Rate (TPR), False Positive Rate (FPR).TPR is the percentage of confidential documents correctly identified, FPR represents the percentage of non-confidential documents mistakenly classified as confidential documents. The final objective of my method is maximizing TPR and minimizes FPR.

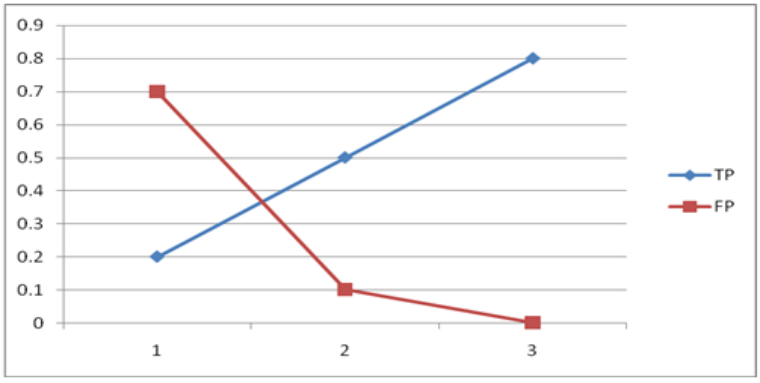

Figure.3. Representation of TP and FP for confidential terms score>200.

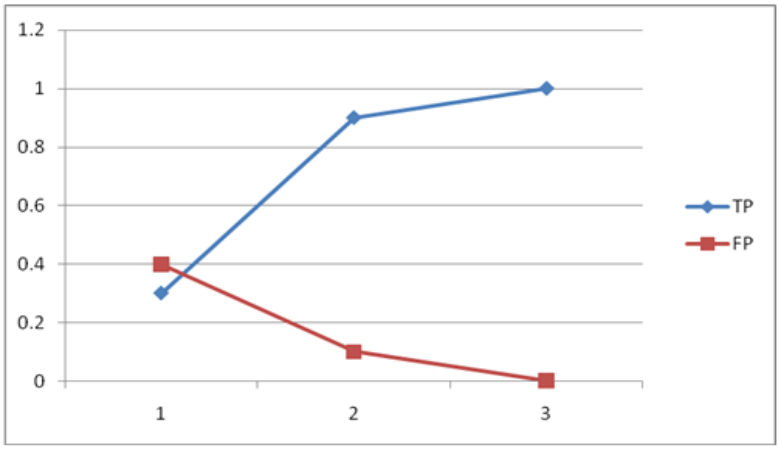

Fig.4. Representation of TP and FP for confidential terms score between 150 and 200

5. CONCLUSION

We have presented a method for data leakage prevention. This method designed for two test cases, detection of entire confidential document and detection of small portions of confidential content embedded in larger non-confidential documents. In both test cases, the presented method maximizes true positive rate and minimizes false positive rate.

REFERENCES

[1] Jinhyung Kim, Jun Hwang & Hyung, (2012) "Privacy level indicating data leakage prevention system", IJSA, Vol 6, No: 3, pp 91-96.

[2] Amir Harel & Asaf Shabtai, (2011)"M-Score: Estimating the potential damage of data leakage incident by assigning misuseability weight", M.Sc Degree.

[3] Data Leakage Prevention Implementation and challenges http://www.niiconsulting.com/innovation/DLP.pdf.

[4] Gilad Katz, Yuval Elovici & Bracha Shapira, (2014)"CoBAn: A Context based model for data leakage prevention", Science Direct, Information Science, pp137-158.

[5] Carvalho & Cohen W. W, (2007)"Preventing information leaks in emails", in proceeding of SIAM International conference on data mining.

[6] Salton G & Buckly (1988)"Term-weighting approaches in automatic text retrieval",Information Processing and Management, pp 513-523.

[7] Gilad Katz, Bracha Shapira & Nir Ofek, (2013) "CoBAn: A context Based Approach for Text Classification, http://www.ise.bgu.ac.il/engineering/upload/23944/technical_report.pdf.

[8] Lavrenko, V & W. B.Croft,(1998)"A language modeling approach to information retrieval", In the proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval ACM, Melbourne,Australia,pp275-281.

[9] Zhai C &Lafferty, (2001)"A study of smoothing methods for language models applied to adhoc information retrieval", in proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, ACM.

[10] Hiemstra D, (2002)"Term-specific smoothing for the language modelling approach to information retrieval: the importance of a query term", in proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval ACM.

[11] https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[12] Information week global security survey. Information Week, 2004.

[13] W.W.Cohen, (1996)"Learning rules that classify e-mail", In Proceedings of the AAAI spring symposium on machine learning in information access, pp18-25.
[14] J.I Helfman, C.I Isbell & Ishmail, (1995)"Immediate identification of important information", AT&T Labs Technical report.
[15] J.D.M & Rennie, (2000),"An application of machine learning to e-mail filtering", in proceedings of the KDD workshop on Text mining.
[16] J.Staddon & P.Golle (2008),"A content-driven access control system", in proceedings of the 7th symposium on identity and trust on the Internet, ACM, Gaithersburg, Marylands, pp26-35.
[17] W.W Cohen & Y.Singer, (1999)"Context-sensitive learning methods for text categorization", ACM transactions on Information sytermmms, pp141-173.
[18] H.Drucker & D.Wu, (1999)"Support vector machines for spam categorization", IEEE transaction on neural networks.
[19] I.Androutsopoulos & J.Koutsias,(2000),"An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with person e-mail messages", In proceedings of the 23rd annual international ACM SIGR conference on Research and Development on Information Retrieval,ACM.Athens,Greece,pp160-167.
[20] M.Sahami & S.Dumais, (1998)"A Bayesian approach to filtering junk email", AAAI-98 workshop on Learning for text categorization.
[21] J.Hovold, (2005),"Naive Bayes spam filtering using word-position-based attributes", in proceedings of the 2nd conference on Email and Anti-spam.
[22] G.Salton, (1983)"Introduction to modern Information retrieval", McGraw Hill, New York, pp448.
[23] J.Song & H.Takakura, (2013)"Toward more practical unsupervised anomaly detection System", Information Science, pp4-14.

AUTHORS

Dr.B.Padmaja Rani has received B.Tech (ECE) from Osmania University, M. Tech (CS) from JNTU, Hyderabad and PhD from JNTU, Hyderabad. She is currently working as Professor in Department of CSE, JNTU, Hyderabad. She is having 15+ years of professional experience and 10+ years of research experience. Her area of interest includes Information Retrieval, Natural Language Processing, Information Security and Cloud Computing.

Subhashini Peneti has received B.Tech (IT) from JNTU, Hyderabad, M.Tech (SE) from JNTU, Hyderabad. She is currently pursuing PhD in JNTU, Hyderabad.